STAT 730 Chapter 4: Estimation

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Analysis

The likelihood

We have *iid* data, at least initially. Each datum comes from a pdf or pmf indexed by $\boldsymbol{\theta}$:

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} f(\mathbf{x}_i; \boldsymbol{\theta}).$$

The likelihood of $\boldsymbol{\theta}$ is simply the joint distribution of $\mathbf{X}$, as a function of $\boldsymbol{\theta}$:

$$L(\mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}).$$

The log-likelihood is the log of the likelihood:

$$l(\mathbf{X}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\theta}).$$

Log-likelihood of multivariate normal data

Note that

$$
\begin{aligned}
\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}) \\
&= \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + 0 \\
&= \mathrm{tr}\left\{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right\} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
&= \mathrm{tr}\left\{\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\right\} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
&= \mathrm{tr}\{n\boldsymbol{\Sigma}^{-1}\mathbf{S}\} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).
\end{aligned}
$$

So

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

implies

$$l(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

Matrix differentiation

Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$. Then $\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}$ is the $n \times p$ matrix with ijth entry $\frac{\partial f(\mathbf{X})}{\partial x_{ij}}$.

If $\mathbf{x} \in \mathbb{R}^n$ is a vector, then $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \in \mathbb{R}^n$ is called the gradient. The (symmetric) matrix of second partials $\mathbf{H} = \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right]$ is called the Hessian.

If $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}) \cdots h_q(\mathbf{x})] \in \mathbb{R}^{1\times q}$ then $\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}}$ is the $p \times q$ matrix with ijth element $\frac{\partial h_i(\mathbf{x})}{\partial x_j}$.

Score function

In general, the score function is

$$\mathbf{s}(\mathbf{X};\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} l(\mathbf{X};\boldsymbol{\theta}) = \frac{1}{L(\mathbf{X};\boldsymbol{\theta})}\frac{\partial}{\partial\theta} L(\mathbf{X};\boldsymbol{\theta}).$$

Note that if $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ then $\mathbf{s} \in \mathbb{R}^p$.

As a function of $\mathbf{X}$, $\mathbf{s}$ is random. $V(\mathbf{s}) = \mathbf{F}$ is called the Fisher information matrix.

Expectation of $\mathbf{s}$

thm: Let $\mathbf{t} \in \mathbb{R}^q$ be a function of $\mathbf{X}$ and $\boldsymbol{\theta}$. Then under some regularity conditions

$$E(\mathbf{s}\mathbf{t}') = \frac{\partial}{\partial \boldsymbol{\theta}} E(\mathbf{t}') - E\left(\frac{\partial \mathbf{t}'}{\partial \boldsymbol{\theta}}\right).$$

$\boxed{\text{Proof}}$: By definition $E\{\mathbf{t}(\mathbf{X};\boldsymbol{\theta})'\} = \int \mathbf{t}(\mathbf{X};\boldsymbol{\theta})' L(\mathbf{X};\boldsymbol{\theta}) d\mathbf{X}$. Differentiate both sides, right side using product rule, subtract off first portion of right-hand side:

$$\frac{\partial E\{\mathbf{t}(\mathbf{X};\boldsymbol{\theta})'\}}{\partial \boldsymbol{\theta}} = \int \left[\frac{\partial \mathbf{t}(\mathbf{X};\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}} L(\mathbf{X};\boldsymbol{\theta}) + \underbrace{\frac{\partial L(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}_{\mathbf{s}(\mathbf{X};\boldsymbol{\theta}) L(\mathbf{X};\boldsymbol{\theta})} \mathbf{t}(\mathbf{X};\boldsymbol{\theta})' \right] d\mathbf{X}. \square$$

Note that $E(\mathbf{s}\mathbf{t}') \in \mathbb{R}^{p \times q}$.

Corollaries

Corollary: $E(\mathbf{s}) = \mathbf{0}$.

$\boxed{\text{Proof}}$: Let $\mathbf{t} = [1]$. $\square$

Corollary: Let $\mathbf{t} = \mathbf{t}(\mathbf{X})$ only and $E(\mathbf{t}) = \boldsymbol{\theta}$ then $E(\mathbf{st}') = \mathcal{I}_p$.

$\boxed{\text{Proof}}$: $\frac{\partial \mathbf{t}'}{\boldsymbol{\theta}} = \mathbf{0}$. $\square$

Corollary: $\mathbf{F} = V(\mathbf{s}) = -E\left(\frac{\partial \mathbf{s}'}{\partial \boldsymbol{\theta}}\right) = -E\left(\left[\frac{\partial^2 \log L(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]\right)$.

Information

The Fisher information $\mathbf{F}$ is the expected matrix of negative 2nd partials of $\log L(\mathbf{X}; \boldsymbol{\theta})$. It has information on the average curvature of $L(\mathbf{X}; \boldsymbol{\theta})$ at $\boldsymbol{\theta}$.

For example, if $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known, then $\mathbf{F} = \left[\frac{n}{\sigma^2}\right]$. The larger this is, the more "peaked" $L(\mathbf{X}; \boldsymbol{\theta})$ is at $\hat{\mu} = \bar{x}$. This happens when either n gets large or σ gets small.

Intuitively, when σ gets small there is more information for each piece of data for μ, so the curvature increases.

Maximization result

<u>thm</u>: Let $\mathbf{A}(p \times p)$ and $\mathbf{B} > 0$ be symmetric. The maximum (minimum) of $\mathbf{x}'\mathbf{A}\mathbf{x}$ given $\mathbf{x}'\mathbf{B}\mathbf{x} = 1$ is given when $\mathbf{x}$ is the e-vector corresponding to the largest (smallest) e-value of $\mathbf{B}^{-1}\mathbf{A}$. That is, $\max_{\mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_1$ and $\min_{\mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_p$ where $\lambda_1 \geq \cdots \geq \lambda_p$ are e-values of $\mathbf{B}^{-1}\mathbf{A}$.

$\boxed{\text{Proof}}$: Let $\mathbf{y} = \mathbf{B}^{1/2}\mathbf{x}$. Want $\max_{\mathbf{y}} \mathbf{y}'\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{y}$ subject to $\mathbf{y}'\mathbf{y} = 1$. Now take $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$ and $\mathbf{z} = \boldsymbol{\Gamma}'\mathbf{y}$. Then $\mathbf{z}'\mathbf{z} = \mathbf{y}'\mathbf{y}$ and we want $\max_{\mathbf{z}} \mathbf{z}'\boldsymbol{\Lambda}\mathbf{z} = \sum_{i=1}^{p} \lambda_i z_i^2$ subject to $\mathbf{z}'\mathbf{z} = 1$. Then we have $\max \sum_{i=1}^{p} \lambda_i z_i^2 \leq \lambda_1 \sum_{i=1}^{p} z_i^2 = \lambda_1$ and this bound is attained when $\mathbf{z} = (1, 0, \ldots, 0)'$, $\mathbf{y} = \gamma_{(1)}$, and $\mathbf{x} = \mathbf{B}^{-1/2}\gamma_{(1)}$. $\mathbf{B}^{-1}\mathbf{A}$ and $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ have the same e-values and $\mathbf{x} = \tilde{\gamma}_{(1)} = \mathbf{B}^{-1/2}\gamma_{(1)}$ is the e-vector of $\mathbf{B}^{-1}\mathbf{A}$ corresponding to λ_1. Minimization proceeds similarly. $\square$

Maximization result, continued

<u>lemma</u>: Let $\mathbf{a} \in \mathbb{R}^p$ s.t. $\mathbf{a} \neq \mathbf{0}$. Then $||\mathbf{a}||^2$ is the only nonzero e-value of $\mathbf{aa}'$ with corresponding e-vector $\frac{\mathbf{a}}{||\mathbf{a}||}$. We will show this in class.

<u>Corollary</u>: For $\mathbf{x}'\mathbf{Bx} = 1$, $\max_{\mathbf{x}} \mathbf{a}'\mathbf{x} = \sqrt{\mathbf{a}'\mathbf{B}^{-1}\mathbf{a}}$ and $\max_{\mathbf{x}}\{(\mathbf{a}'\mathbf{x})^2/(\mathbf{x}'\mathbf{Bx})\} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}$ and the maximum attained at $\mathbf{x} = \mathbf{B}^{-1}\mathbf{a}/\sqrt{\mathbf{a}'\mathbf{B}^{-1}\mathbf{a}}$. $\boxed{\text{Proof}}$: Use $\mathbf{x}'\mathbf{Ax} = \mathbf{x}'[\mathbf{aa}']\mathbf{x}$. $\square$

<u>Corollary</u>: $\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{Aa}}{\mathbf{a}'\mathbf{Ba}} = \lambda_1$ and $\min_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{Aa}}{\mathbf{a}'\mathbf{Ba}} = \lambda_p$ as before, attained at $\mathbf{a} = \gamma_{(1)}$ & $\mathbf{a} = \gamma_{(p)}$ from $\mathbf{B}^{-1}\mathbf{A}$.

$\boxed{\text{Proof}}$: Proceeds exactly as in the theorem. $\square$

Cramér-Rao lower bound

How good can an unbiased estimated of $\boldsymbol{\theta}$ be?

thm: If $\mathbf{t} = \mathbf{t}(\mathbf{X})$ s.t. $E(\mathbf{t}) = \boldsymbol{\theta}$ based on regular likelihood function, then $V(\mathbf{t}) \geq \mathbf{F}^{-1}$.

$\mathbf{A} \geq \mathbf{B} \Leftrightarrow \mathbf{a}'\mathbf{A}\mathbf{a} \geq \mathbf{a}'\mathbf{B}\mathbf{a}$ for all $\mathbf{a}$. Standard covariance result gives $C(\mathbf{a}'\mathbf{t}, \mathbf{c}'\mathbf{s}) = \mathbf{a}'C(\mathbf{t}, \mathbf{s})\mathbf{c} = \mathbf{a}'\mathbf{c}$ (corollary two slides ago) and $V(\mathbf{c}'\mathbf{s}) = \mathbf{c}'V(\mathbf{s})\mathbf{c} = \mathbf{c}'\mathbf{F}\mathbf{c}$. Then

$$\mathrm{corr}^2(\mathbf{a}'\mathbf{t}, \mathbf{c}'\mathbf{s}) = \frac{(\mathbf{a}'\mathbf{c})^2}{\mathbf{a}'V(\mathbf{t})\mathbf{a}\ \mathbf{c}'\mathbf{F}\mathbf{c}} \leq 1.$$

Maximizing this w.r.t. $\mathbf{c}$ subject to $\mathbf{c}'\mathbf{F}\mathbf{c} = 1$ (last slide) gives

$$\frac{\mathbf{a}'\mathbf{F}^{-1}\mathbf{a}}{\mathbf{a}'V(\mathbf{t})\mathbf{a}} \leq 1,$$

for all $\mathbf{a}$. $\square$

Sufficiency

What statistics have all the information for $\boldsymbol{\theta}$?

def'n $\mathbf{t} = \mathbf{t}(\mathbf{X})$ is sufficient for $\boldsymbol{\theta} \Leftrightarrow L(\mathbf{X}; \boldsymbol{\theta}) = g(\mathbf{t}; \boldsymbol{\theta}) h(\mathbf{X})$.

Note that $\mathbf{s}$ depends on $\mathbf{X}$ only through $\mathbf{t}$.

A sufficient statistic is minimal sufficient if it is a function of every other sufficient statistic. Rao-Blackwell (Lehmann-Scheffé elsewhere) theorem says if a minimal sufficient statistic is also complete, then any unbiased estimator that is a function of the minimal sufficient statistic is the unique minimum variance unbiased estimator (MVUE).

Recall: $\mathbf{t}$ complete $\Leftrightarrow E\{g(\mathbf{t})\} = 0$ all $\boldsymbol{\theta} \Rightarrow P_{\boldsymbol{\theta}}\{g(\mathbf{t}) = 0\} = 1$ all $\boldsymbol{\theta}$. Hard to show in general, but exponential families often have complete statistics.

thm: $\bar{\mathbf{x}}$ and $\mathbf{S}$ are complete for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Normal example: sufficiency

For *iid* normal data

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

we have

$$L(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\right\}.$$

So $(\bar{\mathbf{x}}, \mathbf{S})$ are sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; they are also minimally sufficient complete, although the book doesn't discuss this much. So $\bar{\mathbf{x}}$ is MVUE of $\boldsymbol{\mu}$ and $\frac{n}{n-1}\mathbf{S}$ is MVUE of $\boldsymbol{\Sigma}$.

Maximum likelihood estimation

def'n: The MLE $\hat{\boldsymbol{\theta}}$ is $\text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\mathbf{X}; \boldsymbol{\theta})$.

- $\mathbf{s} = \mathbf{0}$ at $\hat{\boldsymbol{\theta}}$. Since $\mathbf{s}$ is a function of a sufficient statistic, so is $\hat{\boldsymbol{\theta}}$. That is, $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\mathbf{t}; \boldsymbol{\theta}) h(\mathbf{X})$, maximized at function of $\mathbf{t}$.
- If $f(\mathbf{x}; \boldsymbol{\theta})$ satisfies regularity conditions then $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{F}^{-1})$ where $\mathbf{F}$ is Fisher information for one observation. This is for *iid* data; a similar result holds for independent but not identically distributed, e.g. regression data.
- This implies $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$ under mild conditions.
- $\hat{\boldsymbol{\theta}}$ is asymptotically unbiased and efficient. Hence the popularity of MLEs. Note that moment-based estimators are also typically asymptotically unbiased but not necessarily efficient.

Minimization result

First note (p. 478) that

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}, \quad \frac{\partial \mathbf{x}'\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}, \quad \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}, \quad \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{y}.$$

Any of these are shown by expanding the forms into sums, taking derivatives, then recognizing the sums as matrix products.

<u>thm</u>: The $\mathbf{x}$ which minimizes $f(\mathbf{x}) = (\mathbf{y} - \mathbf{A}\mathbf{x})'(\mathbf{y} - \mathbf{A}\mathbf{x})$ solves $\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{A}'\mathbf{y}$.

$\boxed{\text{Proof}}$:

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}[\mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{A}'\mathbf{y} + \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x}] = 0 - 2\mathbf{A}'\mathbf{y} + 2\mathbf{A}'\mathbf{A}\mathbf{x}.$$

Set equal to zero and solve. Note that the 2nd derivative matrix $2\mathbf{A}'\mathbf{A} \geq 0$ so sol'n is minimum. $\square$

Maximization result

<u>thm</u> For any $\mathbf{A} > 0$, $f(\mathbf{\Sigma}) = |\mathbf{\Sigma}|^{-n/2} \exp\{-\frac{1}{2}\text{tr } \mathbf{\Sigma}^{-1}\mathbf{A}\}$ is maximized by $\mathbf{\Sigma} = \frac{1}{n}\mathbf{A}$.

$\boxed{\text{Proof}}$: Write $\log f(\frac{1}{n}\mathbf{A}) - \log f(\mathbf{\Sigma}) = \frac{1}{2}np(a - 1 - \log g)$ where $a = \text{tr } \mathbf{\Sigma}^{-1}\mathbf{A}/np$ and $g = |\frac{1}{n}\mathbf{\Sigma}^{-1}\mathbf{A}|^{1/p}$ are the arithmetic and geometric means of the e-values of $\frac{1}{n}\mathbf{\Sigma}^{-1}\mathbf{A}$. All e-values are positive and $a - 1 - \log g \geq 0$ so $f(\frac{1}{n}\mathbf{A}) \geq f(\mathbf{\Sigma})$ for all $\mathbf{\Sigma} > 0$. $\square$

MLEs for normal data: unconstrained

Take $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume $\boldsymbol{\Sigma} > 0$. Recall

$$l(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{n}{2} \text{tr } \boldsymbol{\Sigma}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

First consider $\boldsymbol{\mu}$. As a function of $\boldsymbol{\mu}$, $l(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is maximized (for *any* $\boldsymbol{\Sigma}$) when
$(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = [(\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu})]'[(\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu})]$
is minimized. (Either stare at it or take the first partials w.r.t. $\boldsymbol{\mu}$.)
The minimization result two slides ago implies this occurs when $\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, so $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. It remains to maximize
$L(\mathbf{X}; \hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = c|2\pi\boldsymbol{\Sigma}|^{n/2} \exp\{-\frac{n}{2}\text{tr } \boldsymbol{\Sigma}^{-1}\mathbf{S}\}$, but we have $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ from the last slide.

MLEs for normal data: constrained

If $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is known a priori, $\hat{\boldsymbol{\Sigma}} = \mathbf{S} + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'$ by maximizing
$L(\mathbf{X}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = c|\boldsymbol{\Sigma}|^{-n/2} \exp\{-\frac{n}{2} \text{tr } \boldsymbol{\Sigma}^{-1}[\mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)']\}$.

If $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ is known, $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ as before.

We will use these results in simple hypothesis testing in Chapter 5.

Normal data: MLEs under various constraints

- Know $\boldsymbol{\mu} = \kappa \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is given. Then $\hat{\kappa} = \frac{\boldsymbol{\mu}_0' \mathbf{S}^{-1} \bar{\mathbf{x}}}{\boldsymbol{\mu}_0' \mathbf{S}^{-1} \boldsymbol{\mu}_0}$.
- Know $\mathbf{R}\boldsymbol{\mu} = \mathbf{r}$ (linear constraints) where ($\mathbf{r}, \mathbf{R}$ are given. Then $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \mathbf{S}\mathbf{R}'[\mathbf{R}\mathbf{S}\mathbf{R}']^{-1}(\mathbf{R}\bar{\mathbf{x}} - \mathbf{r})$.

Both of these assume $\boldsymbol{\Sigma}$ unknown; if $\boldsymbol{\Sigma}$ known – which will never happen – replace $\mathbf{S}$ with $\boldsymbol{\Sigma}$ in the above expressions.

- Know $\boldsymbol{\Sigma} = \kappa \boldsymbol{\Sigma}_0$. Then $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\kappa} = \text{tr } \boldsymbol{\Sigma}_0^{-1} \mathbf{S}/p$ (p. 107).
- Know $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, i.e. $\mathbf{x}_{i1}$ indep. $\mathbf{x}_{i2}$ for all $\mathbf{x}_i' = (\mathbf{x}_{i1}', \mathbf{x}_{i2}')$. Then $\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}$.
- If have $\mathbf{X}_i(n_i \times p)$ indep. d.m. from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \ldots, k$, then $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + \cdots + n_k} \sum_{i=1}^k n_i \mathbf{S}_i$.

Bayesian inference

Bayesian inference treats $\boldsymbol{\theta}$ as random and assigns $\boldsymbol{\theta}$ a prior distribution. Inference is then based on the distribution of $\boldsymbol{\theta}$ updated by the data, i.e. the posterior density

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \propto L(\mathbf{X};\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

For normal data

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$\boldsymbol{\mu}$ is typically thought about independently of $\boldsymbol{\Sigma}$ so $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu})p(\boldsymbol{\Sigma})$.

Bayesian inference: Priors on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Common priors for $\boldsymbol{\mu}$ include $\boldsymbol{\mu} \sim N_p(\mathbf{m}, \mathbf{V})$ and the improper flat prior $p(\boldsymbol{\mu}) \propto 1$.

Common priors for $\boldsymbol{\Sigma}$ include $\boldsymbol{\Sigma}^{-1} \sim W_p(\mathbf{A}, a)$ and the improper prior $p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2}$.

The density of $\mathbf{M} \sim W_p(\mathbf{A}, m)$ is given by

$$p(\mathbf{M}) = \frac{|\mathbf{M}|^{(m-p-1)/2} \exp(-\frac{1}{2}\mathrm{tr}\mathbf{A}^{-1}\mathbf{M})}{2^{mp/2}\pi^{p(p-1)/4}|\mathbf{A}|^{m/2} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(m+1-i))}.$$

Bayesian inference: Gibbs sampling

Although it is possible to explicitly obtain the posterior for $\boldsymbol{\mu}|\mathbf{X}$ (it is a multivariate t distribution, p. 110), we shall use a more common approach to obtaining posterior inference, Gibbs sampling.

Gibbs sampling for normal data iteratively samples the two full conditional distributions $[\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathbf{X}]$ and $[\boldsymbol{\Sigma}|\boldsymbol{\mu}, \mathbf{X}]$. Let $\boldsymbol{\mu}^0$ be given. Then the jth iterate is sampled $[\boldsymbol{\Sigma}^j|\boldsymbol{\mu}^{j-1}, \mathbf{X}]$ then $[\boldsymbol{\mu}^j|\boldsymbol{\Sigma}^j, \mathbf{X}]$ for $j = 1, \ldots, J$ where J is some large number. The iterates $\{(\boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j)\}_{j=1}^J$ form a dependent sample from the joint posterior $[\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}]$.

Assume $\boldsymbol{\mu} \sim N_p(\mathbf{m}, \mathbf{V})$ indep. $\boldsymbol{\Sigma}^{-1} \sim W_p(\mathbf{A}, a)$. In your homework you will show

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathbf{X} \sim N_p([n\boldsymbol{\Sigma}^{-1} + \mathbf{V}^{-1}]^{-1}[n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} + \mathbf{V}^{-1}\mathbf{m}], [n\boldsymbol{\Sigma}^{-1} + \mathbf{V}^{-1}]^{-1}),$$

and

$$\boldsymbol{\Sigma}^{-1}|\boldsymbol{\mu}, \mathbf{X} \sim W_p\left(\left[\mathbf{A}^{-1} + \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\right]^{-1}, a + n\right).$$